# Generalized S-Estimators

Christophe Croux, Peter J. Rousseeuw, and Ola Hössjer *

In this paper we introduce a new type of positive-breakdown regression method, called a generalized S-estimator (or *GS-estimator*), based on the minimization of a generalized M-estimator of residual scale. We compare the class of GS-estimators with the usual S-estimators, including least median of squares. It turns out that GS-estimators attain a much higher efficiency than S-estimators, at the cost of a slightly increased worst-case bias. We investigate the breakdown point, the maxbias curve and the influence function of GS-estimators. We also give an algorithm for computing GS-estimators, and apply it to real and simulated data.

KEY WORDS: Breakdown point; Influence function; Maxbias curve; Regression analysis; Robustness.

AMS subject classifications: 62F35, 62J05.

*Christophe Croux is Research Assistant with the Belgian National Fund for Scientific Research, and Peter J. Rousseeuw is Professor, both at the Department of Mathematics and Computer Science, Universitaire Instelling Antwerpen (U.I.A.), Universiteitsplein 1, B-2610 Antwerp, Belgium. Ola Hössjer is Lecturer at the Department of Mathematical Statistics, Lund Institute of Technology, Box 118, S-22100 Lund, Sweden. His work was supported by the Swedish Natural Science Research Council, contract F-DP 6689-300. The authors wish to thank the referees for helpful remarks.

# 1 INTRODUCTION

In the linear model one often has to cope with outliers, which can make the classical least squares (LS) estimator highly unreliable. In fact, even a single outlier can destroy the LS estimate. Many alternative methods have been proposed. Very often used are M- and GM-estimators (see for example Hampel et al 1986), but their breakdown point goes down to zero when the dimension increases. The least median of squares (LMS) and least trimmed squares (LTS) estimators (Rousseeuw 1984) have a 50% breakdown point but a low asymptotic efficiency. A generalization is given by S-estimators (Rousseeuw and Yohai 1984), which can attain an efficiency up to 33% (Hössjer 1992). Both MM-estimators (Yohai 1987) and $\tau$-estimators (Yohai and Zamar, 1988) can attain arbitrarily high efficiency without losing their 50% breakdown point, but they pay for this with an increased bias. Note that MM-estimators need a high-breakdown start, for which we can use one of the estimators discussed below.

In this paper we introduce a new class of regression estimators, called *generalized S-estimators* (or *GS-estimators*), which can have a 50% breakdown point like S-estimators, but attain a much higher efficiency. As a special case of GS-estimators we propose the *least quartile difference* (LQD) estimator, which we define as

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\, Q_n(r_1, \ldots, r_n), \tag{1.1}$$

where $r_i$ is the residual of the $i$ th case, and

$$Q_n = \{|r_i - r_j|;\, i < j\}_{\binom{h_p}{2}:\binom{n}{2}} \tag{1.2}$$

is a scale estimator proposed by Rousseeuw and Croux (1993). Expression (1.2) means that $Q_n$ is the $\binom{h_p}{2}$-th order statistic among the $\binom{n}{2}$ elements of the set $\{|r_i - r_j|;\, i < j\}$. Here, $h_p = [(n + p + 1)/2]$ where $p$ is the number of regression parameters. It turns out that the objective function (1.2) can be computed very quickly by using an efficient algorithm. Therefore, the LQD can easily be implemented by adapting an existing LMS program. The gaussian efficiency of the LQD regression is shown to be 67.1%, which is more than twice the efficiency of any S-estimator with 50% breakdown point. Moreover, the LQD does not require the choice of tuning constants.

An important property of GS-estimators is that their objective function does not depend on the intercept term. (The intercept can be estimated afterwards, with high statistical

efficiency and consuming negligible computation time.) Another advantage of GS-estimators is that they are well-suited for models with an asymmetric error distribution, unlike the usual S-estimators of which the objective function only depends on $r_i$ through $|r_i|$, hence positive residuals are attached the same importance as negative residuals of the same size. Therefore, GS-estimators are more generally applicable. Note that the pairwise differences $r_i - r_j$ have a symmetric distribution even when the $r_i$ themselves do not.

# 2 ROBUSTNESS AT FINITE SAMPLES

We will work with a linear model denoted as

$$y_i = \boldsymbol{\beta}_1 x_{i1} + \ldots + \boldsymbol{\beta}_{p-1} x_{i,p-1} + \alpha + \text{error}_i \quad \text{for } i = 1, \ldots, n. \tag{2.1}$$

The parameter to be estimated is $\boldsymbol{\theta} = (\boldsymbol{\beta}, \alpha) \in I\!R^p$, where $\boldsymbol{\beta} \in I\!R^{p-1}$ is the slope and $\alpha$ is the intercept. Our observations are of the form $\mathbf{z}_i = (\mathbf{x}_i, y_i) = (\mathbf{u}_i, 1, y_i) \in I\!R^{p+1}$. This means that the actual explanatory variables are combined in a vector $\mathbf{u}_i \in I\!R^{p-1}$. (We require of course that $p \geq 2$.) As usual, we assume that $n/p > 5$ to avoid the curse of dimensionality.

We define a generalized S-estimator $\hat{\boldsymbol{\beta}}$ as

$$\hat{\boldsymbol{\beta}} = \operatorname*{argmin}_{\boldsymbol{\beta}} s_n(\boldsymbol{\beta}), \tag{2.2}$$

where $s_n(\boldsymbol{\beta})$ is based on the residuals $r_i = y_i - \boldsymbol{\beta}^t \mathbf{u}_i - \alpha$ through the equation

$$\binom{n}{2}^{-1} \sum_{i<j} \rho\left(\frac{r_i - r_j}{s_n(\boldsymbol{\beta})}\right) = k_{n,p}. \tag{2.3}$$

To avoid having multiple solutions (or no solutions) it is better to define

$$s_n(\boldsymbol{\beta}) = \sup\{s > 0;\ \binom{n}{2}^{-1} \sum_{i<j} \rho\left(\frac{r_i - r_j}{s}\right) \geq k_{n,p}\}. \tag{2.4}$$

Note that $s_n(\boldsymbol{\beta})$ does not depend on $\alpha$ because $r_i - r_j$ doesn't. We will require that:

(R) The function $\rho$ is even, non-decreasing on the positive numbers, and continuous at 0 with $\rho(0) = 0$. There are only a finite number of points where $\rho$ is not continuous or non-differentiable. Furthermore $0 < \rho(\infty) < \infty$, and $\rho(c) = \rho(\infty)$ for some $c > 0$.

We will denote $\lim_{n\to\infty} k_{n,p} = k$.

Because $s_n(\boldsymbol{\beta})$ is independent of the intercept, the latter has to be estimated afterwards. We can estimate $\alpha$ by a location estimate based on the numbers $r_i(\hat{\boldsymbol{\beta}}, 0) = y_i - \hat{\boldsymbol{\beta}}^t \mathbf{u}_i$, for instance by using the median or a more efficient 50% breakdown estimator.

We will pay particular attention to the GS-estimator given by

$$\rho(x) = I(|x| \geq 1) \quad \text{and} \quad k_{n,p} = (\binom{n}{2} - \binom{h_p}{2} + 1)/\binom{n}{2}$$

with $h_p = [(n+p+1)/2]$. Then we have that $s_n(\boldsymbol{\beta}) = Q_n(\boldsymbol{\beta}) = \{|r_i - r_j|; i < j\}_{\binom{h_p}{2}:\binom{n}{2}}$. (One should multiply $Q_n(\boldsymbol{\beta})$ with a certain constant factor to make it consistent as a scale estimator, but that is immaterial to our current goal of estimating $\boldsymbol{\beta}$.) Note that $k = 3/4$. The scale estimator $Q_n$ was discussed in Rousseeuw and Croux (1993). We will denote the corresponding regression estimator (1.1) as LQD, because the objective function is approximately the first quartile of the pairwise differences of the residuals. It is instructive to compare this with the LMS objective function, which is given by

$$\{|r_i|; 1 \leq i \leq n\}_{h_p:n}. \tag{2.5}$$

We will prove that the LQD regression always exists, and has the exact fit property and maximal breakdown point. Using those facts, we will show the existence, exact fit property, and maximal breakdown point for a whole class of generalized S-estimators.

Throughout this section we will assume the property:

(H) no $\binom{h_p}{2}$ of the differences $(\mathbf{u}_i - \mathbf{u}_j, y_i - y_j)$ lie on the same vertical hyperplane in $I\!R^p$.

By "vertical hyperplane" we mean a hyperplane containing $(\mathbf{0}, 0)$ and $(\mathbf{0}, 1)$. Note that this condition is stronger than requiring that there are no $h_p$ observations $(\mathbf{x}_i, y_i)$ lying on a vertical hyperplane in $I\!R^{p+1}$. In Figure 1a we see that no $h_p$ observations $(\mathbf{u}_i, y_i)$ are lying on the same affine hyperplane, but nevertheless condition (H) is not satisfied. If the observations follow a continuous distribution, (H) has probability 1.

**Theorem 1.** *Under condition (H), there always exists a solution to* $\operatorname{argmin}_{\boldsymbol{\beta}} Q_n(\boldsymbol{\beta})$.

(All proofs are given in the appendix.) In order to establish the breakdown point we need another regularity condition.

**Definition:** We say that the differences of the $\mathbf{u}_i$ are in general position if no $\binom{p}{2}$ of the $\mathbf{u}_i - \mathbf{u}_j$ with $i < j$ belong to the same hyperplane in $I\!R^{p-1}$.

If the differences of the $\mathbf{u}_i$ are in general position, then also the $\mathbf{u}_i$ themselves are in general position. The latter means that no $p$ of the $\mathbf{u}_i$ lie on the same affine hyperplane in $I\!R^{p-1}$, which is equivalent to saying that no $p$ of the $\mathbf{x}_i$ lie on the same hyperplane in $I\!R^p$. In Figure 1b we have a situation where neither the $\mathbf{u}_i$ nor their differences are in general position, while in Figure 1c we see that it can happen that the $\mathbf{u}_i$ themselves are in general position but their differences are not.

Note that this condition on the differences $\mathbf{u}_i - \mathbf{u}_j$ is more stringent than the condition on the individual $\mathbf{u}_i$, but not much more. If the $\mathbf{u}_i$ have a continuous distribution, both conditions hold almost surely. From a semantic point of view, the $\mathbf{u}_i$ being in "general" position also precludes such linear relations between the $\mathbf{u}_i - \mathbf{u}_j$ (actually, in computational geometry the phrase "general position" is often interpreted in this stronger sense). When a data set contains a few exceptions to general position this does not mean that GS-estimators can no longer be used, but merely that the expression in Theorem 2 below will be slightly reduced.

Let us now look at the finite-sample breakdown point (Donoho and Huber, 1983). The breakdown point of an estimator $T$ at a sample $Z$ is defined as

$$\varepsilon_n^*(T,Z) = \min\{m/n;\ \sup_{Z'} |T(Z) - T(Z')| = \infty\}, \tag{2.6}$$

where $Z'$ is obtained by replacing any $m$ observations by arbitrary points. (We will use for $|.|$ the euclidean norm.)

**Theorem 2.** *If the differences of the* $\mathbf{u}_i$ *are in general position, then the breakdown point of the LQD estimator is given by* $\varepsilon_n^*(\hat{\boldsymbol{\beta}}, Z) = ([(n-p)/2]+1)/n$, *which is the maximal value for any regression equivariant estimator.*

From the proof it follows that we obtain the maximal breakdown point for any objective function $\{|r_i - r_j|;\ i<j\}_{q:\binom{n}{2}}$ where

$$\binom{[(n+p)/2]-1}{2} + p - 1 \le q \le \binom{[(n+p+1)/2]}{2}. \tag{2.7}$$

We have chosen to define the LQD based on the largest rank $\binom{h_p}{2}$ where $h_p = [(n+p+1)/2]$, which is also in accordance with the rank $h_p$ used in the LMS objective (2.5).

From the general relation between the breakdown point and the exact fit property (Rousseeuw and Leroy 1987, page 123) the next result immediately follows.

**Corollary.** *If at least* $h_p = [(n+p+1)/2]$ *of the observations satisfy* $y_i = \boldsymbol{\beta}_0^t \mathbf{u}_i + \alpha_0$ *exactly and the differences of their* $\mathbf{u}_i$ *are in general position, then* $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}_0$ *no matter what the other observations are.*

**Remark 1:** It is easy to see that if we take a high-breakdown estimator $\hat{\alpha}$ for the intercept, with $\varepsilon_n^*(\hat{\alpha}, Y) \geq ([(n-p)/2]+1)/n$ for any univariate sample $Y$, then we also have the maximal breakdown point and the exact fit property for $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}, \hat{\alpha})$.

**Remark 2:** It is not advisable to use GS-estimators for fitting a zero-intercept model to data that were actually generated *with* an intercept, because they estimate the slope of a point cloud regardless of the position of the origin. For instance, let us look at Figure 1d. Applying LQD yields an acceptable slope estimate, indicated by the dotted line, but because of the zero-intercept model the actual LQD regression (solid line) does not fit the data points. Note, however, that a study of the LQD-based residuals, all having the same sign and size, does reveal immediately that a zero-intercept model is inappropriate for these data (model misspecification).

Now we return to GS-estimators of the type (2.2) with general $\rho$-function.

**Theorem 3.** *The existence, the maximal breakdown point and the exact fit property hold for GS-estimators under the same conditions as for the LQD when*

$$\frac{k_{n,p}}{\rho(c)} = \left(\binom{n}{2} - \binom{h_p}{2} + 1\right) \Big/ \binom{n}{2}$$

*which implies that* $k/\rho(c) = 3/4$.

The most popular choice for a $\rho$-function is the biweight, $\rho(x) = \min(3x^2/c^2 - 3x^4/c^4 + x^6/c^6, 1)$. If we choose $c = 0.9958$ (and thus $k = 0.75$) we obtain a 50% breakdown regression estimator which we will call the biweight GS-estimator. This estimator is consistent (Hössjer, Croux and Rousseeuw 1993).

The usual S-estimators, which can be defined by the objective function,

$$s_n = \sup\{s > 0;\ \frac{1}{n}\sum_{i=1}^{n} \rho(\frac{r_i}{s}) \geq k_{n,p}\}, \tag{2.8}$$

have maximal breakdown point if $k_{n,p}/\rho(c) = (n - h_p + 1)/n$. Therefore we obtain the maximal breakdown point if $k/\rho(c) = 1/2$, where $k = \lim_{n\to\infty} k_{n,p}$. The S-estimator with biweight $\rho$-function and 50% breakdown point (hence, $c = 1.547$) will in this paper be called the biweight S-estimator.

# 3 MAXBIAS CURVES

In this section we follow the approach of Martin, Yohai and Zamar (1989). Because the maxbias curve is an asymptotic notion, we first have to determine the functional corresponding to a GS-estimator. We will add to condition (R) that $\rho(\infty) = 1$ (this is only a normalization) and we will drop the condition that "$\rho$ is constant for large $x$". For any distribution $F$ we define the scale functional

$$s(F) = \sup\{s > 0;\, E_F \rho(\frac{r_1 - r_2}{s}) \geq k\}, \tag{3.1}$$

where $r_1$ and $r_2$ are i.i.d. according to $F$ and $0 < k < 1$. We denote by $s(\boldsymbol{\beta}, K)$, where $K$ is the joint distribution of $(\mathbf{u}, y)$, the same scale functional evaluated at the distribution of the residuals $r(\boldsymbol{\beta}) = y - \boldsymbol{\beta}^t \mathbf{u} - \alpha$, where we note that this scale does not depend on $\alpha$. The corresponding GS-estimator then has the functional version $T$ given by $s(T(K), K) = \inf_{\boldsymbol{\beta}} s(\boldsymbol{\beta}, K)$, hence $T$ is regression and affine equivariant.

Suppose that our model distribution $K_0$ of $(\mathbf{u}, y)$ is elliptical about the origin. We may assume w.l.o.g. that $T(K_0) = 0$ due to regression equivariance. Denote by $G_0$ the distribution of the $\mathbf{u}_i$ and by $F_0$ that of the errors. Consider the contamination neighborhood $V_\varepsilon = \{K; K = (1-\varepsilon)K_0 + \varepsilon K^*\}$, where $K^*$ can be any distribution. Then the maxbias curve is given by

$$B_\varepsilon(T) = \sup\{|T(K)|;\, K \in V_\varepsilon\}. \tag{3.2}$$

The asymptotic breakdown point may then be defined as $\varepsilon^* = \inf\{\varepsilon;\, B_\varepsilon(T) = \infty\}$. Suppose that:

(G) $G_0$ is spherical, $P_{G_0}(\mathbf{u}^t \boldsymbol{\beta}) = 0 \quad$ for all $\boldsymbol{\beta} \neq \mathbf{0}$ in $I\!R^p$, and for all $\boldsymbol{\beta}$ the distribution of $\boldsymbol{\beta}^t \mathbf{u}$ is unimodal;

(F) $F_0$ has a unimodal, continuous and symmetric density.

Define the functions

$$g(s, \boldsymbol{\beta}) = E_{K_0 \times K_0} \rho(\frac{y_1 - y_2 - \boldsymbol{\beta}^t(\mathbf{u}_1 - \mathbf{u}_2)}{s}) \quad \text{and} \quad \tilde{g}(s, \boldsymbol{\beta}) = E_{K_0} \rho(\frac{y - \boldsymbol{\beta}^t \mathbf{u}}{s}). \tag{3.3}$$

Since $G_0$ is spherical, $g$ and $\tilde{g}$ only depend on $|\boldsymbol{\beta}|$ and $s$. It holds that $g$ and $\tilde{g}$ are continuous, strictly increasing in $|\boldsymbol{\beta}|$ and strictly decreasing in $s$ (for $s > 0$). This can be proven as in Lemma 3.1 of Martin, Yohai and Zamar (1989) by using the fact that the distributions of $\mathbf{u}_1 - \mathbf{u}_2$ and $y_1 - y_2$ also satisfy (G) and (F). Therefore we may define $g_1^{-1}(\cdot, |\boldsymbol{\beta}|)$ as the inverse

of $g$ w.r.t. $s$, and $g_2^{-1}(\varepsilon, s)$ as the solution $|\boldsymbol{\beta}|$ of $\tilde{h}(\varepsilon, s, |\boldsymbol{\beta}|) = k$, where $k = \lim_{n\to\infty} k_{n,p}$ was defined below (2.4) and

$$\tilde{h}(\varepsilon, s, |\boldsymbol{\beta}|) = (1-\varepsilon)^2 g(s, |\boldsymbol{\beta}|) + 2\varepsilon(1-\varepsilon)\tilde{g}(s, |\boldsymbol{\beta}|) \quad (0 < \varepsilon < 1). \tag{3.4}$$

By means of two lemmas, given in the appendix, we now obtain the maxbias curve:

**Theorem 4.** *Under the conditions (G) and (F) above, we have*

$$\begin{aligned} B_\varepsilon(T) &= g_2^{-1}(\varepsilon, g_1^{-1}(\frac{k - 2\varepsilon + \varepsilon^2}{(1-\varepsilon)^2}, 0)) \quad \textit{for} \quad \varepsilon \le \min(\sqrt{1-k}, 1-\sqrt{1-k}) \\ &= \infty \quad \textit{elsewhere.} \end{aligned}$$

In the case were $(\mathbf{u}, y)$ is multivariate normal $N(\mathbf{0}, \mathbf{I}_p)$ we obtain

$$\tilde{g}(s, \gamma) = h(\frac{\sqrt{1+\gamma^2}}{s}) \text{ and } g(s, \gamma) = h(\frac{\sqrt{2(1+\gamma^2)}}{s}),$$

where $h(\lambda) = E\rho(\lambda u)$ for $u \sim N(0,1)$. Note that $h(\lambda)$ is increasing and continuous for $\lambda > 0$, so that $h^{-1}$ is well-defined. We can compute $B_\varepsilon(T)$ for any $\varepsilon < \varepsilon^*$ by first computing

$$s_1 = \sqrt{2} \Big/ h^{-1}(\frac{k - 2\varepsilon + \varepsilon^2}{(1-\varepsilon)^2}).$$

Then $B_\varepsilon(T)$ is given by the solution of the following equation in $\gamma$:

$$(1-\varepsilon)^2 h(\frac{\sqrt{2(1+\gamma^2)}}{s_1}) + 2\varepsilon(1-\varepsilon)h(\frac{\sqrt{1+\gamma^2}}{s_1}) = k.$$

In the special case where $\rho$ is a step function $\rho(x) = I(|x| > c)$, we find that $h(\lambda) = 2(1 - \Phi(c/\lambda))$ and thus $h^{-1}(t) = c/\Phi^{-1}(1 - t/2)$.

**Remark 1:** Theorem 4 implies that the breakdown point becomes

$$\varepsilon^* = \min(\sqrt{1-k}, 1-\sqrt{1-k}). \tag{3.5}$$

So for the step function $\rho(x) = I(|x| > c)$ with $P_{F_0}(y_1 - y_2 \le c) = 1 - k/2$, yielding the $(1-k)$th quantile of the pairwise differences of residuals, we obtain (3.5). This gives the maximal breakdown point $\varepsilon^* = 50\%$ when $k = 3/4$, which corresponds to the LQD.

**Remark 2:** The derivative of the maxbias curve at 0 is infinite, hence the gross-error sensitivity of GS-estimators is infinite. This is a property of all regression estimators with dimension-free maxbias curve (He and Simpson, 1993). However, as Yohai and Zamar (1992)

have proposed, one could define a modified gross-error-sensitivity as $\gamma^{**} = \lim_{\varepsilon \downarrow 0} B_\varepsilon(T)/\sqrt{\varepsilon}$. In the case of a multivariate normal distribution, we obtain $\gamma^{**} = \left(2\sqrt{2}(1-h(1))\Big/h'(\sqrt{2})\right)^{1/2}$. For step functions $\rho(x) = I(|x| > c)$ we obtain $\gamma^{**} = \left(2\sqrt{2}(2\Phi(c)-1)\Big/(c\phi(c/\sqrt{2}))\right)^{1/2}$ where $c = \sqrt{2}\Phi^{-1}(1-k/2)$. In particular, for $k = 0.75$ we obtain $\gamma^{**}(\text{LQD}) = 2.399$. One can compare this with $\gamma^{**}(\text{LMS}) = 2.160$. For the biweight GS we obtain $\gamma^{**} = 2.412$, and for the biweight S-estimator $\gamma^{**} = 2.267$.

Figure 2a plots the maxbias curve of the LQD estimator, together with those of the LMS and the biweight S- and GS-estimators. We see that the bias of the LQD estimator is only slightly larger than that of the biweight S, whereas we will see that its efficiency is much better (67% instead of 28%). Switching from LQD to the biweight GS, which has a smooth $\rho$ function, again increases the bias only a little (but the efficiency gain will also be small). In Figure 2a we also see the maxbias curve of the TAU67-estimator (Yohai and Zamar, 1988). The latter estimator is based on two biweight $\rho$-functions, the first with $c = 1.547$ to obtain a 50% breakdown point, the second with $c = 3.26$ to obtain a gaussian efficiency of 67.1%. (We use the TAU67 estimator here for comparison with the LQD, which has the same efficiency.) We see that the maxbias curve of TAU67 is very close to that of the LQD. Furthermore $\gamma^{**}$(TAU67)=2.442, thus for small $\varepsilon$ the LQD behaves slightly better. (Note that the maxbias curve of a $\tau$-estimator with 95% efficiency is somewhat higher.)

In view of Remark 1, we note that there are often two different step functions $\rho$ which yield GS-estimators with the same breakdown point. For example, $k_1 = 0.5$ (corresponding with the median interpoint distance) and $k_2 = 1-(1-\sqrt{0.5})^2 \approx 0.91$ both yield a 29% breakdown estimator. It is interesting to see that their maxbias curves behave rather differently (see Figure 2b): the smaller value of $k$ (which corresponds with the higher quantile) is preferable. Note that the maxbias curve of the LQD(0.5)-estimator is very close to the LQD for up to about 25% of contamination. Afterwards the LQD(0.5) bias increases rapidly, while that of the LQD increases more slowly.

A referee asked to compute the breakdown rate (BR), as defined by Mazzi (1991) and Zamar (1992). For a 50% breakdown estimator $T$, the BR is given by

$$\text{BR}(T) = \lim_{\varepsilon \uparrow 0.5} \frac{B_\varepsilon(T)}{B_\varepsilon(LMS)}. \tag{3.6}$$

Following the proofs and computations in Mazzi (1991), we obtain that BR(LQD)=$0.5+\sqrt{2}$ in the gaussian case. The breakdown rate of the biweight S- and GS-estimators, and also of

TAU67, equals infinity. Therefore, the maxbias of the LQD is lower than that of the biweight S and TAU67 in a neighborhood of the breakdown point.

# 4 INFLUENCE FUNCTION AND EFFICIENCY

The influence function (see Hampel et al 1986) describes the (standardized) effect of a single outlier on the estimator. It is an asymptotic notion, based on the same vector-valued functional $T(K)$ as in the previous section. Let $K_0$ be a fixed distribution representing the central model, and let $K_\varepsilon = (1-\varepsilon)K_0 + \varepsilon\Delta_{\mathbf{u},y}$ where $\Delta_{\mathbf{u},y}$ is the distribution which puts all its mass at the point $(\mathbf{u}, y)$. Then the influence function is defined as

$$IF(\mathbf{u}, y) = \lim_{\varepsilon \downarrow 0} \frac{T(K_\varepsilon) - T(K_0)}{\varepsilon}. \tag{4.1}$$

Another definition is given by a von Mises expansion: if there exists a function $IF$: $I\!R^{p-1} \times I\!R \to I\!R^{p-1}$ (which depends on the estimator and underlying distribution $K_0$) such that

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0 - \frac{1}{n}\sum_{i=1}^{n} IF(\mathbf{u}_i, y_i)) = o_P(1), \tag{4.2}$$

then we call $IF$ the influence function. Under regularity conditions both definitions coincide. The latter definition has the advantage that it readily implies that, if $E[IF(\mathbf{u}, y)] = \mathbf{0}$ and $E|IF(\mathbf{u}, y)|^2 < \infty$, the estimator $\hat{\boldsymbol{\beta}}_n$ is consistent and asymptotically normal with asymptotic covariance matrix

$$V = E[IF(\mathbf{u}, y)IF(\mathbf{u}, y)^t]. \tag{4.3}$$

Hössjer, Croux and Rousseeuw (1993) proved the asymptotic normality of GS-estimators, with the function $IF$ specified below, under the conditions:

(G') The distribution of the $\mathbf{u}_i$ satisfies $E_{G_0}[\mathbf{u}] = 0$ and $E_{G_0}|\mathbf{u}|^3 < \infty$, and $E_{G_0}[\mathbf{u}\mathbf{u}^t]$ is positive definite;

(F') The error distribution $F_0$ has a unimodal density $f$, which is twice differentiable with a bounded second derivative.

We may assume (due to equivariance) that $T(K_0) = \boldsymbol{\beta}_0 = 0$ and that $s(\mathbf{0}, K_0) = 1$, in which case the influence function is given by the following theorem:

**Theorem 5.** *If the model distribution $K_0$ satisfies assumptions (G') and (F'), and if $\rho$ satisfies (R), then the influence function of the generalized S-functional is given by*

$$IF(\mathbf{u}, y) = \frac{\overline{\psi}(y)}{E_{F_0}[\overline{\psi}'(y)]}(E_{G_0}[\mathbf{u}\mathbf{u}^t])^{-1}\mathbf{u}, \tag{4.4}$$

$$\textit{where } \overline{\psi}(y) = E_{F_0}[\psi(y - Y)] \textit{ and } \psi = \rho'.$$

From Formula (4.3) we can then compute asymptotic covariances. The efficiency of a GS-estimator at the gaussian model is thus $e = \left(\int \overline{\psi}'(y)d\Phi(y)\right)^2 \Big/ \int \overline{\psi}(y)^2 d\Phi(y)$ . If $f$ is symmetric, the influence function of the LQD estimator becomes

$$IF(\mathbf{u}, y) = -\frac{f(y - c) - f(y + c)}{2\int f'(c + y)f(y)dy}(E_{G_0}[\mathbf{u}\mathbf{u}^t])^{-1}\mathbf{u}, \tag{4.5}$$

where $P_{F_0}(y_1 - y_2 \leq c) = 5/8$. In Figure 3a we made a plot of this influence function when $(\mathbf{u}, y) \sim N(\mathbf{0}, \mathbf{I}_2)$. We see that for fixed $\mathbf{u}$, the influence function is redescending (in fact, it goes exponentially to zero for $y$ tending to infinity). On the other hand, the function $IF(\cdot, y)$ is unbounded in $\mathbf{u}$, hence the overall influence function is not bounded. From (4.5) it follows that the total influence is small when $|y|$ is large, except when $|\mathbf{u}|$ is exponentially large compared to $|y|$. In the latter situation, the influence function is large at a point $(\mathbf{u}, y)$ lying in a direction with a very small inclination, whose effect on the actual estimated slope is therefore negligible.

For the biweight GS-estimator we obtain a very similar plot (see Figure 3b). This illustrates that the LQD estimator, with its non-smooth $\rho$-function, does have an influence function very similar to that obtained with a smooth $\rho$-function. There is also hardly a gain in efficiency: 68.4% for the biweight GS compared to 67.1% for LQD. This is very different from the situation for usual S-estimators, where the quantile objective functions (like LMS) yield estimators converging at a lower rate.

In Figure 3c we see the influence function of the usual biweight S-estimator. It has the same shape as the GS-estimator, but it is steeper. That is the reason why its efficiency is lower (28.7%). In Figure 3d we plotted the influence function of the LTS, which is still steeper and corresponds to an even lower efficiency.

One might argue that the generalized S-estimator should be compared to an S-estimator with $\tilde{\rho}(y) = E\rho(y - Y) - E\rho(Y)$ for its $\rho$-function. This indeed yields an S-estimator with high efficiency, but with a lower breakdown point and a higher maxbias curve.

Formula (4.4) also holds for some unbounded $\rho$ functions. If we take for example $\rho(y) = y^2$ then we obtain the least squares estimator (the objective function is the standard deviation). This estimator is extremely sensitive to outliers in $\mathbf{u}$ and in $y$. If we take $\rho(y) = |y|$ we obtain Gini's average difference as objective function, corresponding to Wilcoxon scores. From Figure 4a we see that this estimator protects against vertical outliers, but not at all against bad leverage points. The latter estimator can be seen as a smoothed version of the Least Absolute Deviations estimator ($L_1$ estimator), which corresponds to a plain S-estimate with the same $\rho$-function. Its influence function is plotted in Figure 4b. Again, the efficiency increases (from 63.6% to 95.5%) when working on the pairwise differences instead of the individual residuals. Finally, in Figure 4c we see the influence function of the optimal robust 95% efficient Mallows estimator (see Hampel et al 1986), which is bounded. Analogously, Figure 4d gives the $IF$ of the 95% efficient Schweppe estimator.

# 5 COMPUTATION AND SIMULATION

In order to compute a GS-estimator we have to minimize the objective function $s(\boldsymbol{\beta})$, where $\boldsymbol{\beta}$ is a p-dimensional vector. There has been a substantial amount of research on algorithms for S-estimators, especially the LMS. The same kind of techniques can be used for computing GS-estimators, including the LQD estimator.

The basic scheme for computing S-estimators is the p-subset algorithm (Rousseeuw and Leroy 1987), which minimizes the objective function over all $\boldsymbol{\beta}_J$ which correspond to fitting a subset $J$ with $p$ observations (out of the $n$ available points). Note that the p-subset version of the LMS is itself a high-breakdown regression estimator (Rousseeuw and Bassett 1991), which is also true for the LQD. Therefore we can use

$$\hat{\boldsymbol{\beta}}^* = \underset{\boldsymbol{\beta}_J}{\operatorname{argmin}}\, Q_n(y_i - \boldsymbol{\beta}_J^t \mathbf{u}_i), \tag{5.1}$$

where $\boldsymbol{\beta}_J$ is determined by the p-subset $J$. If we use the efficient algorithm of Croux and Rousseeuw (1992) to compute $Q_n$ then this objective function merely needs $O(n \log n)$ operations, yielding an overall computation time of $O(n^{p+1} \log n)$ if *all* p-subsets are considered. By comparison, the exhaustive p-subset algorithm for LMS needs $O(n^{p+1})$ time, and also needs $O(n^{p+1} \log n)$ if the intercept is adjusted in every step. Therefore the LQD needs no more computation time than the LMS, while achieving a much better statistical efficiency.

Note that the p-subset algorithm can be modified to run much faster (this holds for all estimators of this type, including LMS, LQD, S- and GS-estimators). The idea is not to consider all $\binom{n}{p} = O(n^p)$ possible p-subsets, but instead to use only $O(n)$ such subsets according to a particular design (Rousseeuw 1993) which ensures that the regression estimator still has the deterministic 50% breakdown point. The resulting LQD algorithm needs only $O(n^2 \log n)$ operations.

Computing the objective function of the biweight GS takes $O(n^2)$ operations (using a fixed number of iterations to solve equation (2.3)), which is more time consuming than the LQD. One can reduce the actual computation time, although it remains $O(n^2)$, in the following way. When considering $m$ trial values $\boldsymbol{\beta}_J$ we don't need to compute $s(\boldsymbol{\beta}_J)$ each time. Indeed, suppose that $\tilde{s}$ is the currently best scale. Generalizing an observation of Yohai and Zamar (1991), we then have

$$s(\boldsymbol{\beta}_J) \leq \tilde{s} \quad \Leftrightarrow \quad \sum_{i<j} \rho(\frac{r_i - r_j}{\tilde{s}}) < k_{n,p} \binom{n}{2}. \tag{5.2}$$

Therefore, we only have to compute a new scale estimate when (5.2) holds. This happens $O(\log m)$ times. At each new best estimate $\tilde{\boldsymbol{\beta}}$ it is possible to carry out some local improvement as in Ruppert (1992). The smoothness of our objective function indicates that Newton steps can be useful. For this, we compute

$$\Delta(\boldsymbol{\beta}) = \frac{s(\boldsymbol{\beta})}{d} \Lambda^{-1} \sum_{i<j} \psi(\frac{r_i - r_j}{s(\boldsymbol{\beta})})(\mathbf{u}_i - \mathbf{u}_j), \tag{5.3}$$

where $\Lambda_{k,l} = \sum_{i<j}(u_{i,l} - u_{j,l})(u_{i,k} - u_{j,k})^t$ and $d = E\psi'(y_1 - y_2) \; = E_\Phi \psi'(y/\sqrt{2})$. We search for the smallest value of $k(\leq 10)$ for which $s(\boldsymbol{\beta} + 2^{-k}\Delta(\beta)) < \tilde{s}$, if there is any. An additional trick is to use

$$(r_i - r_j)(\boldsymbol{\beta} + 2^{-(k+1)}\Delta(\boldsymbol{\beta})) = \frac{1}{2}(r_i - r_j)(\boldsymbol{\beta}) + \frac{1}{2}(r_i - r_j)(\boldsymbol{\beta} + 2^{-k}\Delta(\boldsymbol{\beta})) \tag{5.4}$$

to speed up the computation. Some experiments with this algorithm show that the objective function will be computed only a few times. The number $m$ is obtained by a tradeoff between robustness and speed of computation. When computation time permits, carrying out the Newton steps at each $\boldsymbol{\beta}_J$ is even more accurate.

**Remark:** Stromberg (1993) has given an exact $O(n^{p+2})$ algorithm for the LMS, which can be generalized to the LQD estimator because

$$\min_{\boldsymbol{\beta}} |r_i - r_j|_{\binom{h_p}{2}} = \min_{\boldsymbol{\beta}} \min_{J \in \mathcal{C}_h} \max_{(i,j) \in J} |r_i - r_j| = \min_{J \in \mathcal{C}_h} \min_{\boldsymbol{\beta}} \max_{(i,j) \in J} |r_i - r_j|,$$

where $\mathcal{C}_h = \{J \subset \{1,\ldots,n\}^2;$ for all $(i,j) \in J : i < j$ and $\#J = \binom{h_p}{2}\}$. Thus we have to compute Chebyshev fits $\boldsymbol{\beta}_c$ on $\{(\mathbf{u}_i - \mathbf{u}_j, y_i - y_j);\ (i,j) \in J\}$ for all possible $J$. Only at these values $\boldsymbol{\beta}_c$ do we have to compute the objective function, which needs $O(n \log n)$ operations. Now using a theorem of Cheney (1966, page 36), it is sufficient to look at Chebyshev fits on collections of $p$ observations $(\mathbf{u}_i - \mathbf{u}_j, y_i - y_j)$ with $i < j$. Because the computation time for such a fit only depends on $p$, we obtain a total time of $O(n^{2p+1} \log n)$, which of course is only practical for small values of $n$ and $p$.

The data in Table 1 were obtained from T. Vos of the EPFL in Lausanne (Switzerland). The experiment went as follows. Labeled nitrogen (nitrogen-15) was administered to barley plants in the form of fertilizer ($NH_4NO_3$) in order to study the nitrogen cycle. The nitrogen is taken up by the plants and converted, after a certain time, to organic material in the soil. The purpose of the study was to explain the organic nitrogen by means of other variables. The variables included in the study are: time (in days) after addition of nitrogen ($x_1$), nitrogen content in mineral form in the soil ($x_2$), nitrogen content in the plants ($x_3$), and nitrogen content in organic form in the soil ($y$).

Following Rousseeuw and van Zomeren (1990), we made a diagnostic plot (Figure 5a) of the standardized robust residuals $r_i/Q(r_1,\ldots,r_n)$ obtained by the LQD method, versus robust distances $\mathrm{RD}_i$ obtained with the MVE estimator. In this plot we can identify 6 good leverage points and 2 bad leverage points. The latter (cases 13 and 14) stand out considerably. (Note that also a designed experiment can yield leverage points!) A diagnostic plot of LS residuals versus Mahalanobis distances (see Figure 5b) does not reveal outliers or leverage points. It would be possible to apply LS regression to this data without cases 13 and 14 (while keeping the good leverage points, since they augment the efficiency).

We also performed a small simulation study based on 1000 samples $\{(u_i, y_i);\ i = 1,\ldots,n\}$ from a bivariate gaussian distribution with unit covariance matrix, for various sample sizes $n$. For each sample we computed the LMS, LTS, LQD, S-, GS-, and TAU67 estimators by means of the exhaustive p-subset algorithm. Table 2 lists the resulting finite-sample efficiencies of these estimators, where those of the scale estimators were normalized as in Rousseeuw and Croux (1993).

For the slope, we note that LQD outperforms both LMS and LTS, the gain being larger for increasing $n$. The finite-sample efficiencies of LMS, LTS, and LQD all converge quite

Table 1: Nitrogen Data Set with Robust Distances of $(x_{i1}, x_{i2}, x_{i3})$ based on the MVE, as well as Standardized Residuals $r_i/\hat{\sigma}$ from the LQD Regression

| $i$ | $x_{i1}$ | $x_{i2}$ | $x_{i3}$ | $y_i$ | $\mathrm{RD}_i$ | $r_i/\hat{\sigma}$ |
|---|---|---|---|---|---|---|
| 1 | 0.00 | 61.45 | 0.00 | 12.18 | 0.72 | 0.46 |
| 2 | 0.00 | 58.11 | 0.01 | 6.57 | 0.63 | -0.50 |
| 3 | 0.00 | 65.35 | 0.01 | 6.99 | 0.84 | -0.29 |
| 4 | 1.00 | 47.94 | 0.22 | 10.69 | 0.46 | -0.07 |
| 5 | 1.00 | 57.85 | 0.13 | 13.75 | 0.63 | 0.60 |
| 6 | 1.00 | 35.23 | 0.28 | 10.84 | 0.61 | -0.29 |
| 7 | 4.00 | 44.12 | 0.40 | 15.94 | 0.41 | 0.54 |
| 8 | 4.00 | 33.19 | 0.39 | 9.41 | 0.52 | -0.71 |
| 9 | 4.00 | 24.18 | 0.40 | 17.81 | 0.77 | 0.46 |
| 10 | 19.00 | 25.03 | 2.40 | 17.46 | 0.84 | -0.25 |
| 11 | 19.00 | 30.61 | 3.43 | 23.78 | 0.46 | 0.92 |
| 12 | 19.00 | 23.28 | 3.67 | 18.84 | 0.84 | 0.00 |
| 13 | 49.00 | 2.76 | 29.67 | 57.08 | 23.00 | 5.43 |
| 14 | 49.00 | 1.87 | 26.75 | 48.11 | 19.88 | 3.84 |
| 15 | 49.00 | 1.04 | 23.59 | 23.26 | 16.52 | -0.29 |
| 16 | 80.00 | 0.87 | 26.08 | 35.37 | 13.38 | 0.17 |
| 17 | 80.00 | 0.44 | 31.00 | 28.82 | 18.64 | -0.64 |
| 18 | 80.00 | 0.20 | 23.92 | 30.45 | 11.08 | -0.73 |
| 19 | 111.00 | 0.42 | 18.99 | 44.63 | 0.84 | -0.29 |
| 20 | 111.00 | 0.63 | 23.73 | 51.75 | 5.14 | 1.08 |
| 21 | 111.00 | 0.38 | 22.02 | 51.21 | 3.35 | 0.90 |

Table 2: Finite-Sample Efficiencies of the LMS, LTS, LQD, Biweight S, Biweight GS, and TAU67 Estimators

| | slope | | | | | | scale | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $n$ | LMS | LTS | LQD | S | GS | $\tau$ | LMS | LTS | LQD | S | GS | $\tau$ |
| 10 | 20.8 | 23.1 | 30.2 | 28.1 | 35.2 | 37.2 | 36.1 | 35.0 | 45.3 | 42.5 | 50.0 | 59.2 |
| 20 | 19.9 | 20.0 | 30.7 | 26.9 | 36.5 | 50.8 | 36.6 | 33.0 | 54.2 | 45.7 | 57.9 | 72.2 |
| 40 | 16.9 | 14.9 | 36.0 | 25.5 | 43.5 | 54.3 | 38.5 | 34.0 | 65.2 | 49.2 | 69.4 | 73.8 |
| 60 | 16.0 | 13.9 | 36.8 | 28.1 | 47.0 | 60.5 | 37.2 | 32.5 | 68.5 | 48.4 | 71.3 | 83.5 |
| 80 | 15.3 | 12.8 | 36.9 | 28.3 | 52.2 | 63.2 | 40.2 | 33.1 | 75.2 | 51.7 | 77.8 | 83.5 |
| 100 | 13.4 | 13.4 | 38.6 | 26.8 | 52.1 | 63.9 | 38.2 | 32.1 | 72.1 | 49.4 | 74.1 | 80.9 |
| 200 | 12.8 | 11.6 | 45.3 | 28.5 | 58.5 | 66.4 | 38.7 | 31.9 | 77.0 | 50.4 | 78.3 | 81.0 |
| $\infty$ | 00.0 | 07.1 | 67.1 | 28.7 | 68.4 | 67.1 | 36.7 | 30.7 | 82.3 | 53.9 | 82.9 | 82.7 |

slowly to their asymptotic limits. (Also note that the LMS is more efficient that the LTS for a large range of sample sizes!) The finite-sample efficiencies of the biweight S are quite stable, but they are below those of LQD. The GS- and TAU67 estimators have the best performance overall.

For the corresponding estimators of the error scale we see that the efficiencies of LMS and LTS are rather stable, whereas the others converge more slowly. Also here the LQD, GS- and TAU67 estimators outperform the plain S-estimator, both asymptotically and for finite samples.

# 6 OUTLOOK

Similar to generalized S-estimators, we may construct other classes of high-breakdown estimators. For example, we can define a generalized R-estimator (or *GR-estimator*) as $\hat{\boldsymbol{\beta}}_n = \text{argmin}_{\boldsymbol{\beta}} D_n(r_1, \ldots, r_n)$, where

$$D_n(r_1, \ldots, r_n) = \sum_{i<j} a(R^+(r_i - r_j))|r_i - r_j|. \tag{6.1}$$

Here, $R^+(r_i - r_j)$ stands for the rank of $r_i - r_j$ among the $\binom{n}{2}$ differences $\{r_i - r_j;\, i < j\}$. We assume that the scores are generated by a function $h^+ : [0,1] \to I\!R_+$ using

$$a(i) = \int_{(i-1)/\binom{n}{2}}^{i/\binom{n}{2}} h^+(t)dt.$$

(When $h^+ = 1$ we obtain Wilcoxon scores.) Since the objective (6.1) is location invariant, we can estimate the intercept afterwards.

If $h^+(u) = 0$ for all $u > 1/4$ and $h^+(1/4) > 0$ we obtain a 50% breakdown regression estimator. For instance, if $h^+ = \delta_{1/4}$ we obtain the LQD. When $h^+(u) = I(|u| \leq 1/4)$ we obtain the estimator

$$\hat{\boldsymbol{\beta}}_n = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \sum_{k=1}^{\binom{h_p}{2}} \{|r_i - r_j|;\, i < j\}_{k:\binom{n}{2}}. \tag{6.2}$$

Note that this estimator cannot be written as a GS-estimator. An advantage of GR-estimators is that their objective function (6.1) is explicit, so one does not have to solve an equation. But in most cases (6.1) requires $O(n^2)$ computation time. We think that the maximal efficiency of a GR-estimator wouldn't be much higher than that of the LQD (in fact, the efficiency of (6.2) is 66.04%) and that the LQD can be seen as a prototype of this class of estimators.

If we put $h(t) = h^+(2t-1)$ for $t \in [\frac{1}{2}, 1]$ and $h(t) = -h^+(1-2t)$ for $t \in [0, \frac{1}{2}]$, then the influence function at the model distribution $F$ will be given by

$$IF(\mathbf{u}, y) = \frac{E_F h(\tilde{F}(y - Y))}{B(h, F)} (E[\mathbf{u}\mathbf{u}^t])^{-1}\mathbf{u}, \tag{6.3}$$

where $\tilde{F}$ is the distribution of $y_1 - y_2$ when the $y_i$ are i.i.d. according to $F$, and $B(h, F) = -\int h(\tilde{F}(y))\tilde{F}''(y)dy$. If we further denote $A(h, F) = \int \left(E_F h(\tilde{F}(y - Y))\right)^2 dF(y)$ then we obtain the asymptotic normality $n^{1/2}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \to N(\mathbf{0}, (E[\mathbf{u}\mathbf{u}^t])^{-1} A(h, F) / B^2(h, F))$ .

Instead of working with GS-estimators based on a kernel $\xi(r_i, r_j) = |r_i - r_j|$ of order two, one could also use higher order kernels. If we use a generalized M-estimator (Serfling 1984) as objective function, then $s(\boldsymbol{\beta})$ is defined as the solution of the equation

$$\binom{n}{l}^{-1} \sum_{i_1 < \ldots < i_l} \rho\left(\frac{\xi(r_{i_1}, \ldots, r_{i_l})}{s(\boldsymbol{\beta})}\right) = k_{n,p}.$$

We want $\xi$ to be scale equivariant and location invariant. If we take $\xi(r_{i_1}, \ldots, r_{i_l}) = sdv(r_{i_1}, \ldots, r_{i_l})$, where $sdv$ stands for the standard deviation, we obtain a 50% breakdown estimator if $k/\rho(\infty) = 1 - 2^{-l}$. A prototype of this class is

$$\hat{\boldsymbol{\beta}}_n = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \{sdv(r_{i_1}, \ldots, r_{i_l});\, i_1 < \ldots < i_l\}_{\binom{h_p}{l}:\binom{n}{l}}.$$

These higher order estimators will have a higher efficiency, but also a higher sensitivity to gross errors. We do not recommend these estimators in practice, since their computation time becomes too large.

Looking in a different direction, we note that GS-estimators can also be extended to high-breakdown estimation of scatter matrices. Consider a data set $\mathbf{x}_1, \ldots, \mathbf{x}_n$ of points in $I\!R^p$. Then a GS-estimator of its scatter can be defined as a symmetric positive definite matrix $C$ which minimizes $det(C)$ subject to

$$\binom{n}{2}^{-1} \sum_{i<j} \rho(\| \mathbf{x}_i - \mathbf{x}_j \|_C) \geq \tilde{k}_{n,p} \tag{6.4}$$

where $\| \mathbf{x}_i - \mathbf{x}_j \|_C$ stands for $((\mathbf{x}_i - \mathbf{x}_j)^t C^{-1} (\mathbf{x}_i - \mathbf{x}_j))^{1/2}$. Note that the constraint (6.4) is location invariant, unlike the usual S-estimators (Rousseeuw and Leroy 1987, page 263) where it is necessary to estimate a location vector $T$ simultaneously. A special case of (6.4) is the constraint

$$\{\| \mathbf{x}_i - \mathbf{x}_j \|_C \, ; i < j\}_{\binom{\tilde{h}}{2}:\binom{n}{2}} \geq \tilde{k}, \tag{6.5}$$

yielding an analog to LQD regression. For the computation of the scatter matrices given by (6.4) and (6.5) one can adapt existing algorithms for the MVE and S-estimators.

# 7 APPENDIX

*Proof of Theorem 1.* Denote $M = \max_{i<j} |y_i - y_j|$. Due to condition (H), we have

$$\inf_{|\boldsymbol{\beta}|=1} \{|(\mathbf{u}_i - \mathbf{u}_j)^t \boldsymbol{\beta}|; \, i < j\}_{\binom{h_p}{2}:\binom{n}{2}} = \delta > 0.$$

For $|\boldsymbol{\beta}| > 2M/\delta$ we obtain $|r_i - r_j| \geq ||(\mathbf{u}_i - \mathbf{u}_j)^t \boldsymbol{\beta}| - |y_i - y_j|| \geq |(\mathbf{u}_i - \mathbf{u}_j)^t \boldsymbol{\beta}| - M \geq \delta |\boldsymbol{\beta}| - M > M$ for at least $\binom{n}{2} - \binom{h_p}{2} + 1$ differences $|r_i - r_j|$. Thus for all $|\boldsymbol{\beta}| > 2M/\delta$ it is true that $Q_n(\boldsymbol{\beta}) > M \geq Q_n(\mathbf{0})$. Since $\boldsymbol{\beta} \to Q_n(\boldsymbol{\beta})$ is continuous, $Q_n(\boldsymbol{\beta})$ will attain its minimum value inside the compact ball $\mathcal{B}(\mathbf{0}, 2M/\delta)$. □

*Proof of Theorem 2.* Denote $\varepsilon_n^* = \varepsilon_n^*(\mathrm{LQD}, Z)$. For any regression equivariant estimator we have that $\varepsilon_n^* \leq ([(n-p)/2] + 1)/n$ (Rousseeuw and Leroy 1987, page 125), so it is sufficient to prove the reverse inequality. We can assume w.l.o.g. that $T(Z) = 0$. Denote by $\{(\mathbf{u}_i', y_i'); \, i = 1, \ldots, n\}$ the contaminated sample obtained by replacing $k = [(n-p)/2]$ observations from $Z$, and by $\boldsymbol{\beta}_1$ the corresponding estimate. Denote $M = \max_i |y_i|$, and $r_i' = y_i' - \boldsymbol{\beta}^t \mathbf{u}_i'$. We will prove that $|\boldsymbol{\beta}_1| < C$, where $C$ only depends upon the original sample.

Note that $|r_i(0) - r_j(0)| = |y_i - y_j| \leq 2M$ for all "good" points $(\mathbf{u}_i, y_i)$. Since $\binom{n-k}{2} = \binom{[(n+p+1)/2]}{2} \geq \binom{h_p}{2}$, we have that $Q_n(\mathbf{0}) \leq 2M$. So it is sufficient to prove that for all $|\boldsymbol{\beta}| \geq C$ it holds that $Q_n(\boldsymbol{\beta}) > 2M$, because then it is clear that $|\boldsymbol{\beta}_1| \leq C$. Define

$$\tau = \frac{1}{2}\inf\{\mu > 0; \text{ there is a } (p-2) \text{ dimensional subspace } V \text{ in } (y=0) \text{ such that } V^\mu$$

$$\text{contains } \binom{p}{2} \text{ differences } \mathbf{u}_i - \mathbf{u}_j \ (i<j)\},$$

where $V^\mu$ consists of the points with distance to $V$ less than or equal to $\mu$. Due to our condition of general position, we have $\tau > 0$. Denote $\rho = \tau/n$, and take $C = 10M/\rho$. Take $|\boldsymbol{\beta}| \geq C$ and denote by $H$ the hyperplane in $I\!R^p$ with equation $y = \boldsymbol{\beta}^t\mathbf{u}$. Then $L = H \cap (y=0)$ has dimension $(p-2)$ in $I\!R^{p-1}$.

We will partition the good observations into classes, induced by the following equivalence relation:

$(\mathbf{u}_i, y_i) \sim (\mathbf{u}_j, y_j) \Leftrightarrow$ there exist $k$ $(0 \leq k \leq n-2)$ different observations $\mathbf{u}_{i_1}, \ldots, \mathbf{u}_{i_k}$ such that $\mathbf{u}_i - \mathbf{u}_{i_1} \in L^\rho, \mathbf{u}_{i_2} - \mathbf{u}_{i_3} \in L^\rho, \ldots, \mathbf{u}_{i_k} - \mathbf{u}_j \in L^\rho$.

Denote by $B_1, \ldots, B_m$ the classes with more than one element, and by $B_0$ the union of the other classes. One can see that

$$\mathbf{u}_i, \mathbf{u}_j \in B_l \Rightarrow \mathbf{u}_i - \mathbf{u}_j \in L^\tau \quad \text{for } l \geq 1.$$

Due to the definition of $\tau$, we have $\sum_{l=1}^m \binom{\#B_l}{2} < \binom{p}{2}$.

Now we will divide the "bad" points into subclasses. Denote by $C_l$ $(1 \leq l \leq m)$ the collection of bad points $(\mathbf{u}'_j, y'_j)$ for which there exists an element $(\mathbf{u}_i, y_i)$ in $B_l$ such that $|r'_i - r'_j| \leq (\rho|\boldsymbol{\beta}| - 2M)/4$. For each element $(\mathbf{u}_i, y_i)$ in $B_0$, we denote the collection of bad points $(\mathbf{u}'_j, y'_j)$ for which $|r'_i - r'_j| \leq (\rho|\boldsymbol{\beta}| - 2M)/4$ by $C_{i+m}$ $(1 \leq i \leq m' = \#B_0)$. Note that some of the $C_{i+m}$ can be empty. The bad points that do not belong to any $C_l$ $(1 \leq l \leq m+m')$ are put in $C_0$.

If $\mathbf{u} \notin L^\rho$ then $|\mathbf{u}^t\boldsymbol{\beta}| > \rho|\boldsymbol{\beta}|$. Therefore, if two good observations $(\mathbf{u}_i, y_i)$ and $(\mathbf{u}_j, y_j)$ do not belong to the same class $B_l$ $(l \geq 0)$ or both belong to $B_0$ (and thus $\mathbf{u}_i - \mathbf{u}_j \notin L^\rho$), then

$$|r'_i - r'_j| = |(y_i - y_j) - \boldsymbol{\beta}^t(\mathbf{u}_i - \mathbf{u}_j)| \geq ||y_i - y_j| - |\boldsymbol{\beta}^t(\mathbf{u}_i - \mathbf{u}_j)|| > \rho|\boldsymbol{\beta}| - 2M,$$

where we have used that $|\boldsymbol{\beta}| \geq 2M/\rho$ and $|y_i - y_j| \leq 2M$. Now we can see that the collections $C_l$ $(0 \leq l \leq m+m')$ are disjoint (using the triangular inequality).

From the above it follows that $|r_i' - r_j'| \leq (\rho|\boldsymbol{\beta}| - 2M)/4$ for at most

$$\sum_{l=1}^{m} \binom{\#B_l}{2} + \sum_{l=1}^{m} (\#B_l)(\#C_l) + \sum_{l=1}^{m'} (\#C_{l+m}) + \binom{\sum_{l=0}^{m+m'} \#C_l}{2}$$

couples $i < j$. Using the fact the each $B_l$ $(l \geq 1)$ contains at most $p-1$ elements, we find that the above expression is less than or equal to

$$\binom{p}{2} - 1 + k(p-1) + \binom{k}{2} = \binom{k+p-1}{2} + p - 2,$$

hence $|r_i' - r_j'| > (\rho|\boldsymbol{\beta}| - 2M)/4 > 2M$ at least

$$\binom{n}{2} - \binom{[(n+p)/2] - 1}{2} - p + 2$$

times. Since, if we assume the natural condition $n > p$,

$$\binom{[(n+p)/2] - 1}{2} + (p-1) \leq \binom{[(n+p)/2] - 1}{2} + ([(n+p)/2] - 1) = \binom{[(n+p)/2]}{2} \leq \binom{h_p}{2}$$

we have that $Q_n(\boldsymbol{\beta}) > 2M$. □

*Proof of Theorem 3.* Looking at the proofs of the preceding propositions for the special case where $\hat{\boldsymbol{\beta}}$ is the LQD, it is sufficient to prove that there exist constants $\gamma > 0$ and $\delta > 0$ such that

$$\gamma Q(r_1, \ldots, r_n) \leq s(r_1, \ldots, r_n) \leq \delta Q(r_1, \ldots, r_n), \tag{7.1}$$

where $Q(r_1, \ldots, r_n) = \{|r_i - r_j|; i < j\}_{\binom{h_p}{2}:\binom{n}{2}}$.

We can take $\gamma = 1/c$. Indeed, suppose that $Q > cs$. Then there will be $\binom{n}{2} - \binom{h_p}{2} + 1$ differences $|r_i - r_j|$ greater than $cs$. For an $\varepsilon > 0$ small enough, we will have that

$$\binom{n}{2}^{-1} \sum_{i<j} \rho\left(\frac{r_i - r_j}{s + \varepsilon}\right) \geq \binom{n}{2}^{-1} \left(\binom{n}{2} - \binom{h_p}{2} + 1\right)\rho(c) = k_{n,p}$$

but then $s$ does not satisfy (2.4) any more.

For $\delta$ we can take $\delta = 1/\rho^{-1}(\rho(c)/(\binom{h_p}{2} + 1))$. (We define $\rho^{-1}(u) = \sup\{t > 0;\ \rho(t) \leq u\}$, and then we have that $\rho(\rho^{-1}(u)) \leq u$ for all $u$.) It holds that $\rho^{-1}(\rho(c)/(\binom{h_p}{2} + 1)) > 0$ because otherwise $\rho(t) > \rho(c)/(\binom{h_p}{2} + 1)$ for all $t > 0$. Due to the continuity of $\rho$ at zero this would yield $\rho(c) = 0$, which is a contradiction. Now suppose $s\,\rho^{-1}(\rho(c)/(\binom{h_p}{2} + 1)) > Q$, then $\binom{h_p}{2}$ differences $|r_i - r_j|/s$ are less than $\rho^{-1}(\rho(c)/(\binom{h_p}{2} + 1))$ and thus, for an $\varepsilon$ small enough,

$$\binom{n}{2}^{-1} \sum_{i<j} \rho\left(\frac{r_i - r_j}{s - \varepsilon}\right) \leq \binom{n}{2}^{-1} \rho(c)\left(\binom{h_p}{2}/\left(\binom{h_p}{2} + 1\right) + \binom{n}{2} - \binom{h_p}{2}\right) < k_{n,p}$$

which is in contradiction with (2.4). □

In order to prove Theorem 4, we need the following two lemmas:

**Lemma 1.** *Let* $(\mathbf{u}_1, y_1) \sim K_0$ *and* $(\mathbf{u}_2, y_2) \sim K^*$, *where* $K^*$ *can be any distribution. Then for all* $s > 0$ *and for all* $\boldsymbol{\beta} > 0$*:*

$$E_{K_0 \times K^*}\left[\rho(\frac{y_1 - y_2 - \boldsymbol{\beta}^t(\mathbf{u}_1 - \mathbf{u}_2)}{s})\right] \geq E_{K_0}\left[\rho(\frac{y_1 - \boldsymbol{\beta}^t x_1}{s})\right]. \tag{7.2}$$

*Proof of Lemma 1.* Using symmetry and unimodality of $y_1$ and $\boldsymbol{\beta}^t\mathbf{u}_1$ we find that $z = (y_1 - \boldsymbol{\beta}^t\mathbf{u}_1)/s$ is unimodal and symmetric. Therefore, $E\rho(z - z^*) \geq E\rho(z)$ where $z^*$ can be any stochastic variable. This proves (7.2). □

**Lemma 2.** *Let* $(\mathbf{u}, y) \sim K_0, \mathbf{u}_n^*$ *be uniformly distributed on the line segment* $[\lambda_n\boldsymbol{\beta}^*, 2\lambda_n\boldsymbol{\beta}^*]$, *and put* $y_n = \mathbf{u}_n^t\boldsymbol{\beta}^*$ *and* $K_n^* \sim (\mathbf{u}_n^*, y_n^*)$. *Suppose* $\lambda_n \to \infty$, $\boldsymbol{\beta}_n \to \tilde{\boldsymbol{\beta}}$ *and* $|\tilde{\boldsymbol{\beta}}| < |\boldsymbol{\beta}^*|$. *Then*

$$\lim_{n\to\infty} E_{K_0 \times K_n^*}\left[\rho(\frac{y_1 - y_2 - {\beta_n}^t(\mathbf{u}_1 - \mathbf{u}_2)}{s})\right] = 1 \quad \textit{for all } s > 0 \tag{7.3}$$

$$\lim_{n\to\infty} E_{K_n^* \times K_n^*}\left[\rho(\frac{y_1 - y_2 - {\beta_n}^t(\mathbf{u}_1 - \mathbf{u}_2)}{s})\right] = 1 \quad \textit{for all } s > 0. \tag{7.4}$$

*Proof of Lemma 2.* Take $\varepsilon > 0$. We have that

$$\begin{aligned}
&|E_{K_0 \times K_n^*}\left[\rho(\frac{y_1 - y_2 - {\boldsymbol{\beta}_n}^t(\mathbf{u}_1 - \mathbf{u}_2)}{s})\right] - 1| \\
&\quad = |\frac{1}{\lambda_n}\int_{\lambda_n}^{2\lambda_n} E_{K_0}\left[\rho(\frac{y - {\boldsymbol{\beta}_n}^t\mathbf{u} - t(|\boldsymbol{\beta}^*|^2 - {\boldsymbol{\beta}_n}^t\boldsymbol{\beta}^*)}{s})\right] d\,t - 1| \\
&\quad = |E_{K_0}\left[\rho(\frac{y - {\boldsymbol{\beta}_n}^t\mathbf{u} - t_n(|\boldsymbol{\beta}^*|^2 - {\boldsymbol{\beta}_n}^t\boldsymbol{\beta}^*)}{s})\right] - 1|
\end{aligned} \tag{7.5}$$

where $\lambda_n \leq t_n \leq 2\lambda_n$ (in the last step we used the mean value theorem for integrals of positive continuous functions). Denote $H_n$ the distribution of $(y - {\boldsymbol{\beta}_n}^t\mathbf{u})/s$; since $\boldsymbol{\beta}_n$ is a bounded sequence we can find a compact set $C$ for which $H_n\{C\} > \sqrt{1-\varepsilon}$. Denote further $x_n = t_n(|\boldsymbol{\beta}^*|^2 - {\boldsymbol{\beta}_n}^t\boldsymbol{\beta}^*)/s$, then $x_n \to \infty$ (using that $\lim_{n\to\infty}|\boldsymbol{\beta}^*|^2 - {\boldsymbol{\beta}_n}^t\boldsymbol{\beta}^* > 0$). So for $n$ large enough, we have for all $x \in C$ that $\rho(x - x_n) > \sqrt{1-\varepsilon}$. We can rewrite (7.5) as

$$\begin{aligned}
1 - \int_C \rho(x - x_n)d\,H_n(x) - \int_{\mathbb{R}^p \setminus C} \rho(x - x_n)dH_n(x) &\leq 1 - \int_C \rho(x - x_n)dH_n(x) \\
\leq 1 - \inf_{x\in C} \rho(x - x_n)\,H_n\{C\} &\leq 1 - (1 - \varepsilon) = \varepsilon.
\end{aligned}$$

This proves (7.3). We continue with the proof of (7.4):

$$\begin{aligned}
&|E_{K_n^*\times K_n^*}\left[\rho(\frac{y_1-y_2-{\boldsymbol\beta_n}^t(\mathbf{u}_1-\mathbf{u}_2)}{s})\right]-1| \\
&= |E_{K_n^*\times K_n^*}\left[\rho((\frac{\boldsymbol\beta^*-\boldsymbol\beta_n}{s})^t(\mathbf{u}_1-\mathbf{u}_2))\right]-1| \\
&= |\frac{2}{\lambda_n}\int_0^{\lambda_n}(1-\frac{t}{\lambda_n})\rho((\frac{\boldsymbol\beta^*-\boldsymbol\beta_n}{s})^t\boldsymbol\beta^* t)d\,t-1|, \qquad (7.6)
\end{aligned}$$

where we worked out the distribution of $\mathbf{u}_1-\mathbf{u}_2$. Denote $x_n=(|\boldsymbol\beta^*|^2-{\boldsymbol\beta_n}^t\boldsymbol\beta^*)/s$. Then $x_n>\delta>0$ for $n$ large. Choose $L$ such that $\rho(x_n t)\geq\rho(t\delta)>1-\varepsilon/4$ for all $t>L$. For $n$ big enough we have that $\lambda_n>L$ and $2L/\lambda_n<\varepsilon/2$. Then (7.6) equals

$$\begin{aligned}
&|\frac{2}{\lambda_n}\int_0^L(1-\frac{t}{\lambda_n})(\rho(x_n t)-1)d\,t+\frac{2}{\lambda_n}\int_L^{\lambda_n}(1-\frac{t}{\lambda_n})(\rho(x_n t)-1)d\,t| \\
&\leq \frac{2L}{\lambda_n}+\frac{2}{\lambda_n}(\lambda_n-L)\sup_{t>L}|\rho(x_n t)-1|\leq\frac{2L}{\lambda_n}(1-\varepsilon/4)+\varepsilon/2 \\
&\leq \frac{2L}{\lambda_n}+\varepsilon/2\leq\varepsilon/2+\varepsilon/2=\varepsilon.
\end{aligned}$$

This proves (7.4). □

*Proof of Theorem 4.* Let $c=g_2^{-1}(\varepsilon,s_1)$ where $s_1=g_1^{-1}((k-2\varepsilon+\varepsilon^2)/(1-\varepsilon)^2,0)$. Suppose that $\varepsilon<\min(1-\sqrt{1-k},\sqrt{1-k})$.

We will first prove that $B_\varepsilon(T)\leq c$. Take any distribution $K$ of the form $K=(1-\varepsilon)K_0+\varepsilon K^*$ and consider a slope $|\boldsymbol\beta|>c$. It is sufficient to prove that

$$s(\boldsymbol\beta,K)>s(\mathbf{0},K). \qquad (7.7)$$

Since $\tilde h(\varepsilon,s_1,c)=k$, we have that $\tilde h(\varepsilon,s_1,|\boldsymbol\beta|)>k$. Using continuity, there must exist an $s_2>s_1$ such that $\tilde h(\varepsilon,s_2,|\boldsymbol\beta|)>k$. Using this last inequality, Lemma 1 and the positivity of $\rho$, we find

$$\begin{aligned}
&E_{K\times K}\left[\rho(\frac{y_1-y_2-\boldsymbol\beta^t(\mathbf{u}_1-\mathbf{u}_2)}{s_2})\right] \\
&= (1-\varepsilon)^2 g(s_2,\boldsymbol\beta)+2\varepsilon(1-\varepsilon)E_{K_0\times K^*}\left[\rho(\frac{y_1-y_2-\boldsymbol\beta^t(\mathbf{u}_1-\mathbf{u}_2)}{s_2})\right] \\
&\quad +\varepsilon^2 E_{K^*\times K^*}\left[\rho(\frac{y_1-y_2-\boldsymbol\beta^t(\mathbf{u}_1-\mathbf{u}_2)}{s_2})\right] \\
&\geq (1-\varepsilon)^2 g(s_2,\boldsymbol\beta)+2\varepsilon(1-\varepsilon)\tilde g(s_2,\boldsymbol\beta)=\tilde h(\varepsilon,s_2,\boldsymbol\beta)>k.
\end{aligned}$$

Therefore,

$$s_2\leq s(\boldsymbol\beta,K). \qquad (7.8)$$

Now for any $s > s_1$ we have that

$$\begin{aligned} E_{K\times K}\left[\rho(\frac{y_1-y_2}{s})\right] &\leq (1-\varepsilon)^2 g(s,0) + 2\varepsilon(1-\varepsilon) + \varepsilon^2 \\ &\leq (1-\varepsilon)^2 g(s_1,0) + 2\varepsilon(1-\varepsilon) + \varepsilon^2 = k. \end{aligned}$$

We can conclude that $s \geq s(\mathbf{0}, K)$, and thus

$$s_1 \geq s(\mathbf{0}, K). \tag{7.9}$$

Combining (7.8) and (7.9) yields (7.7).

Now we will prove the other inequality

$$B_\varepsilon(T) \geq c. \tag{7.10}$$

Take any $0 < c_1 < c$ and $|\boldsymbol{\beta}^*| = c_1$. Let $K_n^*$ be a contaminating distribution correponding to $(\mathbf{u}_n^*, y_n^*)$, where $y_n^* = \boldsymbol{\beta}^{*t}\mathbf{u}_n^*$ and $\mathbf{u}_n^*$ is uniformly distributed on the line segment $[\lambda_n\boldsymbol{\beta}^*, 2\lambda_n\boldsymbol{\beta}^*]$ and $\lambda_n \to \infty$. (In fact, we want both location and spread of $\mathbf{u}_n^*$ to go beyond all bounds when $n$ increases.) To prove (7.10), it is enough to show that

$$\sup_n |T(K_n)| \geq c_1, \tag{7.11}$$

where $K_n = (1-\varepsilon) + \varepsilon K_n^*$. Suppose that (7.11) is not true, then we are able to construct a subsequence, which we shall still call $K_n$, for which $\lim_{n\to\infty} \boldsymbol{\beta}_n = \tilde{\boldsymbol{\beta}}$, where $T(K_n) = \boldsymbol{\beta}_n$ and $|\tilde{\boldsymbol{\beta}}| < |\boldsymbol{\beta}^*| = c_1$. Now we have

$$E_{K_n\times K_n}\left[\rho(\frac{y_1-y_2-{\boldsymbol{\beta}_n}^t(\mathbf{u}_1-\mathbf{u}_2)}{s})\right] = (1-\varepsilon)^2 g(s,|\boldsymbol{\beta}_n|)$$

$$+2\varepsilon(1-\varepsilon)E_{K_0\times K_n^*}\left[\rho(\frac{y_1-y_2-{\boldsymbol{\beta}_n}^t(\mathbf{u}_1-\mathbf{u}_2)}{s})\right] + \varepsilon^2 E_{K_n^*\times K_n^*}\left[\rho(\frac{y_1-y_2-{\boldsymbol{\beta}_n}^t(\mathbf{u}_1-\mathbf{u}_2)}{s})\right].$$

Using Lemma 2 and the definition of $s_1$ yields for all $s < s_1$ that

$$\begin{aligned} \lim_{n\to\infty} E_{K_n\times K_n}\left[\rho(\frac{y_1-y_2-{\boldsymbol{\beta}_n}^t(\mathbf{u}_1-\mathbf{u}_2)}{s})\right] &\geq (1-\varepsilon)^2 g(s,0) + 2\varepsilon(1-\varepsilon) + \varepsilon^2 \\ &> (1-\varepsilon)^2 g(s_1,0) + 2\varepsilon(1-\varepsilon) + \varepsilon^2 = k. \end{aligned}$$

Therefore, $\lim_{n\to\infty} s(\boldsymbol{\beta}_n, K_n) \geq s$ and thus

$$\lim_{n\to\infty} s(\boldsymbol{\beta}_n, K_n) \geq s_1. \tag{7.12}$$

On the other hand $\tilde{h}(\varepsilon, s_1, c_1) < \tilde{h}(\varepsilon, s_1, c) = k$. Due to the continuity of $\tilde{h}$ we can find an $s_2 < s_1$ such that $\tilde{h}(\varepsilon, s_2, c_1) < k$. Using the fact that $y_n^* = \boldsymbol{\beta}^{*t}\mathbf{u}_n^*$ exactly, we have that

$$\begin{aligned} & E_{K_n \times K_n}\left[\rho(\frac{y_1 - y_2 - \boldsymbol{\beta}^{*t}(\mathbf{u}_1 - \mathbf{u}_2)}{s_2})\right] \\ & \quad = \quad (1-\varepsilon)^2 g(s_2, c_1) + 2\varepsilon(1-\varepsilon)\tilde{g}(s_2, c_1) = \tilde{h}(\varepsilon, s_2, c_1) < k. \end{aligned}$$

Therefore,

$$s(\boldsymbol{\beta}^*, K_n) \leq s_2. \tag{7.13}$$

Combining (7.12) and (7.13) shows that for $n$ large enough $\boldsymbol{\beta}_n$ does not minimize $s(\cdot, K_n)$, which gives a contradiction. Therefore, (7.11) must be true.

To complete the proof, we show that if $\varepsilon \uparrow \min(\sqrt{1-k}, 1-\sqrt{1-k})$ then

$$g_2^{-1}(\varepsilon, g_1^{-1}(\frac{k - 2\varepsilon + \varepsilon}{(1-\varepsilon)^2}, 0)) \to \infty.$$

This follows from $k - 2\varepsilon + \varepsilon \geq 0 \Leftrightarrow \varepsilon \leq 1 - \sqrt{1-k}$ and $(1-\varepsilon)^2 + 2\varepsilon(1-\varepsilon) \leq k \Leftrightarrow \varepsilon \leq \sqrt{1-k}$.

□

*Proof of Theorem 5.* Combining equations (11), (15) and (17) of Theorem 1 in Hössjer et al. (1993) yields (4.4) according to definition (4.2). Here we will give a proof using definition (4.1). The functional $T$ is given by $T(K) = \text{argmin}_{\boldsymbol{\beta}}\, s(\boldsymbol{\beta}, K)$, where $s(\boldsymbol{\beta}, K)$ satisfies $E_{K \times K}\left[\rho((y_1 - y_2 - \boldsymbol{\beta}^t(\mathbf{u}_1 - \mathbf{u}_2))/s(\boldsymbol{\beta}, K)\right]$, and $(\mathbf{u}_1, y_1)$ and $(\mathbf{u}_2, y_2)$ are two independent variables drawn from $K$. Since $T(K)$ minimizes $s(\boldsymbol{\beta}, K)$ we have

$$E_{K \times K}\left[\psi(\frac{y_1 - y_2 - T(K)^t(\mathbf{u}_1 - \mathbf{u}_2)}{s(T(K), K)})(\mathbf{u}_1 - \mathbf{u}_2)\right] = 0.$$

Also the contaminated distribution $K_\varepsilon = (1-\varepsilon)K_0 + \varepsilon\Delta_{\mathbf{u},y}$ has to satisfy

$$E_{K_\varepsilon \times K_\varepsilon}\left[\psi(\frac{y_1 - y_2 - T(K_\varepsilon)^t(\mathbf{u}_1 - \mathbf{u}_2)}{s(T(K_\varepsilon), K_\varepsilon)})(\mathbf{u}_1 - \mathbf{u}_2)\right] = 0.$$

Working this out yields

$$(1-\varepsilon)^2 E_{K_0 \times K_0}\left[\psi(\frac{y_1 - y_2 - T(K_\varepsilon)^t(\mathbf{u}_1 - \mathbf{u}_2)}{s(T(K_\varepsilon), K_\varepsilon)})(\mathbf{u}_1 - \mathbf{u}_2)\right]$$

$$+2\varepsilon(1-\varepsilon)E_{K_0}\left[\psi(\frac{y_1 - y - T(K_\varepsilon)^t(\mathbf{u}_1 - \mathbf{u})}{s(T(K_\varepsilon), K_\varepsilon)})(\mathbf{u}_1 - \mathbf{u})\right] = 0.$$

Differentiating with respect to $\varepsilon$ and evaluating in 0 gives

$$E_{K_0\times K_0}\left[\psi'(y_1-y_2)\{-\sum_{l=1}^{p} IF_l(\mathbf{u},y,K_0)(u_{1,l}-u_{2,l})-(y_1-y_2)\frac{\partial s(T(K_\varepsilon),K_\varepsilon)}{\partial\varepsilon}|_{\varepsilon=0}\}(u_{1,k}-u_{2,k})\right]$$

$$+2E_{K_0}\left[\psi(y_1-y)(u_{1,k}-u_k)\right]=0$$

for all $1\le k\le p-1$. Since $y$ and $\mathbf{u}$ are independent at $K_0$ and $E_{G_0}(u_{1,k}-u_{2,k})=0$ we find

$$-E_{F_0\times F_0}\left[\psi'(y_1-y_2)\right]IF(\mathbf{u},y)^t E_{G_0\times G_0}\left[(\mathbf{u}_1-\mathbf{u}_2)(u_{1,k}-u_{2,k})\right]$$

$$+2E_{F_0}\left[\psi(y_1-y)\right]E\left[u_{1,k}-u_k\right]=0$$

for all $1\le k\le p-1$. Since $E(\mathbf{u}_1)=0$ and $E_{G_0\times G_0}\left[(\mathbf{u}_1-\mathbf{u}_2)(\mathbf{u}_1-\mathbf{u}_2)^t\right]=2E_{G_0}[\mathbf{u}_1\mathbf{u}_1^t]$ we obtain equation (4.4). □

# REFERENCES


Cheney, E.W. (1966), *Introduction to Approximation Theory,* New York: McGraw-Hill.

Croux, C., and Rousseeuw, P.J. (1992), "Time-Efficient Algorithms for two Highly Robust Estimators of Scale," in *Computational Statistics, Volume 1,* eds. Y. Dodge and J. Whittaker, Heidelberg: Physika-Verlag, 411-428.

Donoho, D.L., and Huber, P.J. (1983), "The Notion of Breakdown Point," in *A Festschrift for Erich Lehmann,* eds. P. Bickel, K. Doksum, and J.L. Hodges, Jr., Wadsworth, California.

Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J., and Stahel, W.A. (1986), *Robust Statistics: the Approach based on Influence Functions,* New York: John Wiley.

He, X., and Simpson, D.G. (1993), "Lower Bounds for Contamination Bias: Globally Minimax versus Locally Linear Estimation," *The Annals of Statistics,* 21, 314–337.

Hössjer, O. (1992), "On the Optimality of S-Estimators," *Statistics and Probability Letters,* 14, 413–419.

Hössjer, O., Croux, C., and Rousseeuw, P.J. (1993), "Asymptotics of Generalized S-Estimators," submitted for publication.

Martin, R.D., Yohai, V.J., and Zamar, R.H. (1989), "Min-Max Bias Robust Regression," *The Annals of Statistics,* 17, 1608–1630.

Mazzi, S.V. (1991), "A New Measure of Quantitative Robustness," Master's Thesis, University of British Columbia, Vancouver.

Rousseeuw, P.J. (1984), "Least Median of Squares Regression," *Journal of the American Statistical Association,* 79, 871–880.

Rousseeuw, P.J. (1993), "A Resampling Design for Computing High-Breakdown Regression," *Statistics and Probability Letters,* 18, 125–128.

Rousseeuw, P.J., and Bassett, G.W. (1991), " Robustness of the $p$-Subset Algorithm for Regression with High Breakdown Point", in *Directions in Robust Statistics and Diagnostics, Part II*, eds. W. Stahel and S. Weisberg, New York: Springer Verlag, 185-194.

Rousseeuw, P.J., and Croux, C. (1993), "Alternatives to the Median Absolute Deviation," *Journal of the American Statistical Association,* 88, 1273-1283.

Rousseeuw, P.J., and Leroy, A.M. (1987), *Robust Regression and Outlier Detection,* New York: John Wiley.

Rousseeuw, P.J., and Yohai, V.J. (1984), "Robust Regression by means of S-estimators," in *Robust and Nonlinear Time Series Analysis,* eds. J. Franke, W. Härdle and R.D. Martin, Lecture Notes in Statistics 26, New York: Springer Verlag.

Rousseeuw, P.J., and van Zomeren, B.C. (1990), "Unmasking Multivariate Outliers and Leverage Points," *Journal of the American Statistical Association,* 85, 633-639.

Ruppert, D. (1992), "Computing S-Estimators for Regression and Multivariate Location/Dispersion," *Journal of Computational and Graphical Statistics*, 1, 253-270.

Serfling, R.J. (1984), "Generalized L-, M-, and R-Statistics," *The Annals of Statistics,* 12, 76–86.

Stromberg, A.J. (1993), "Computing the Exact Least Median of Squares Estimate and Stability Diagnostics in Multiple Linear Regression," *SIAM Journal of Scientific and Statistical Computing,* 14, November issue.

Yohai, V.J. (1987), "High Breakdown Point and High Efficiency Robust Estimates for Regression," *The Annals of Statistics,* 15, 642–656.

Yohai, V.J., and Zamar, R.H. (1988), "High Breakdown Point Estimates of Regression by Means of the Minimization of an Efficient Scale," *Journal of the American Statistical Association,* 83, 406–413.

Yohai, V.J., and Zamar, R.H. (1991), "Discussion of 'Least Median of Squares Estimation in Power Systems' by Mili, L., Phaniraj, V., and Rousseeuw, P.J.," *IEEE Transactions on Power Systems,* 6, 520.

Yohai, V.J., and Zamar, R.H. (1992), "Optimally Bounding a Generalized Gross-Error Sensitivity," manuscript.

Zamar, R.H. (1992), "A New Measure of Quantitative Robustness," presented at the *Workshop on Data Analysis and Robustness,* Ascona (Switzerland), June 29-July 3.

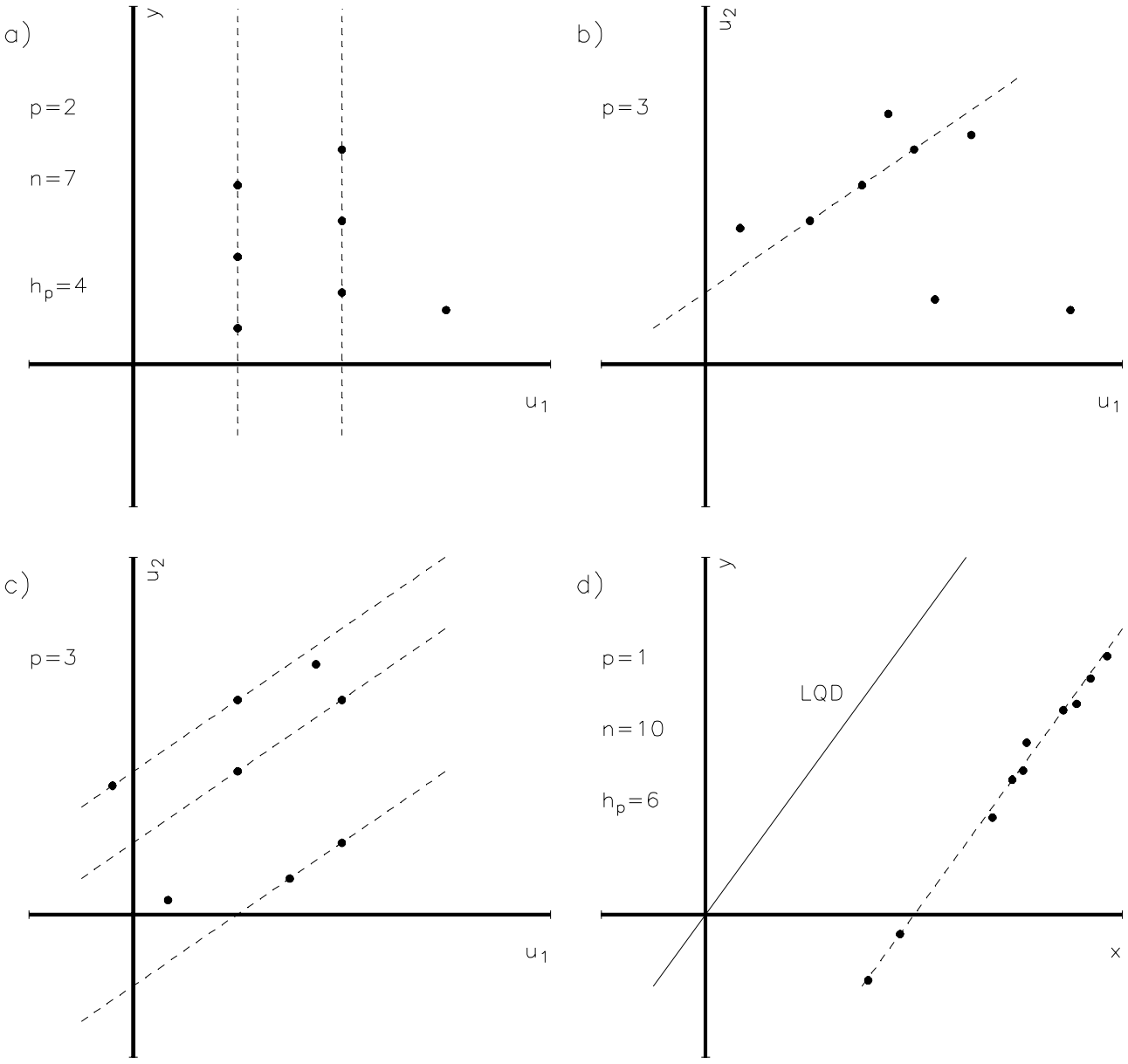


*Figure 1. Examples where: a) condition (H) is not satisfied; b) neither the* $\mathbf{u}_i$ *nor their differences are in general position; c) the* $\mathbf{u}_i$ *are in general position but their differences are not; d) a zero-intercept model is inappropriate.*

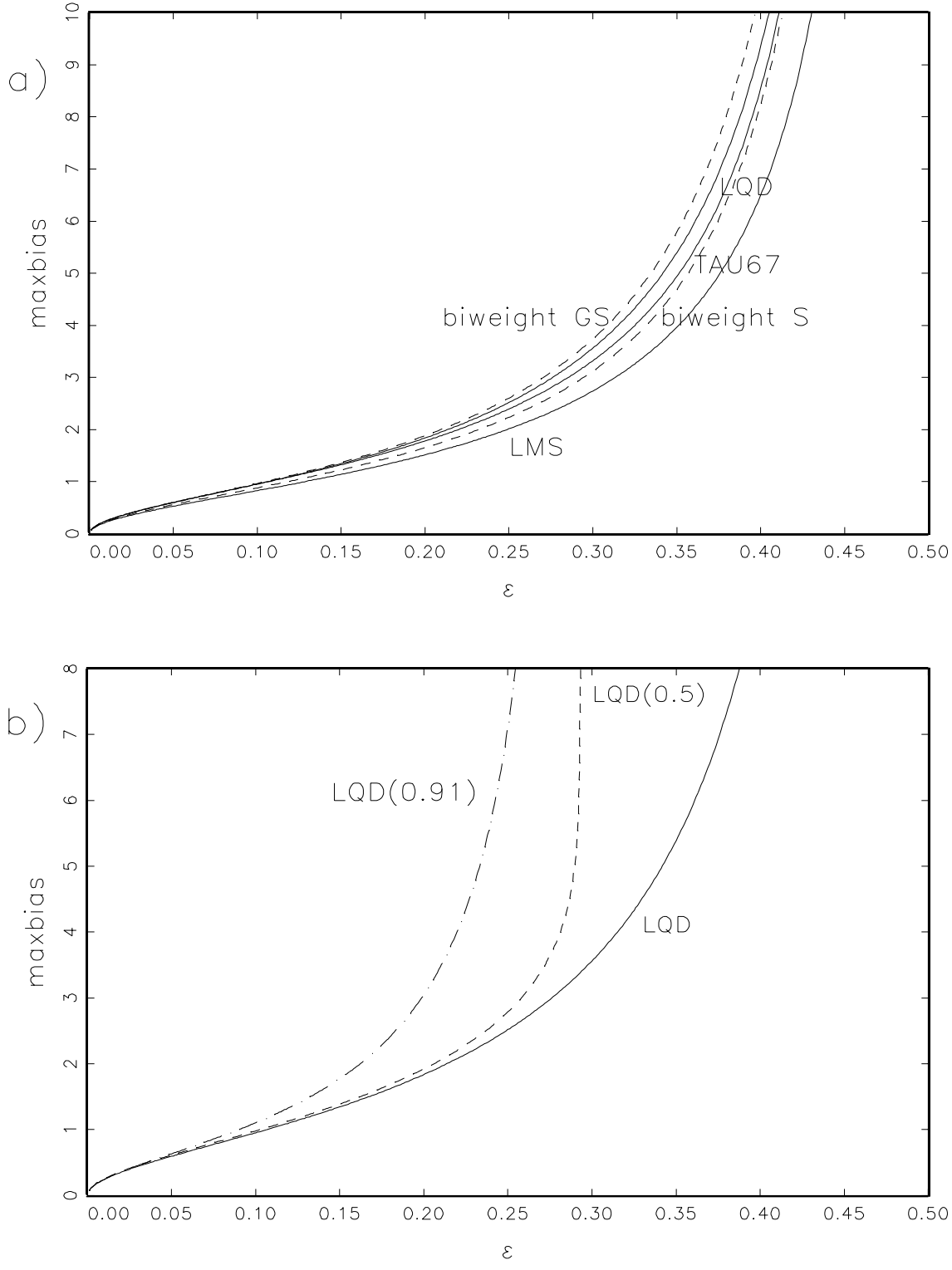





*Figure 2. Maxbias curves of: a) the LQD, LMS, biweight S-, biweight GS-, and the TAU67 estimators; b) the LQD(0.5) and LQD(0.91) estimators.*

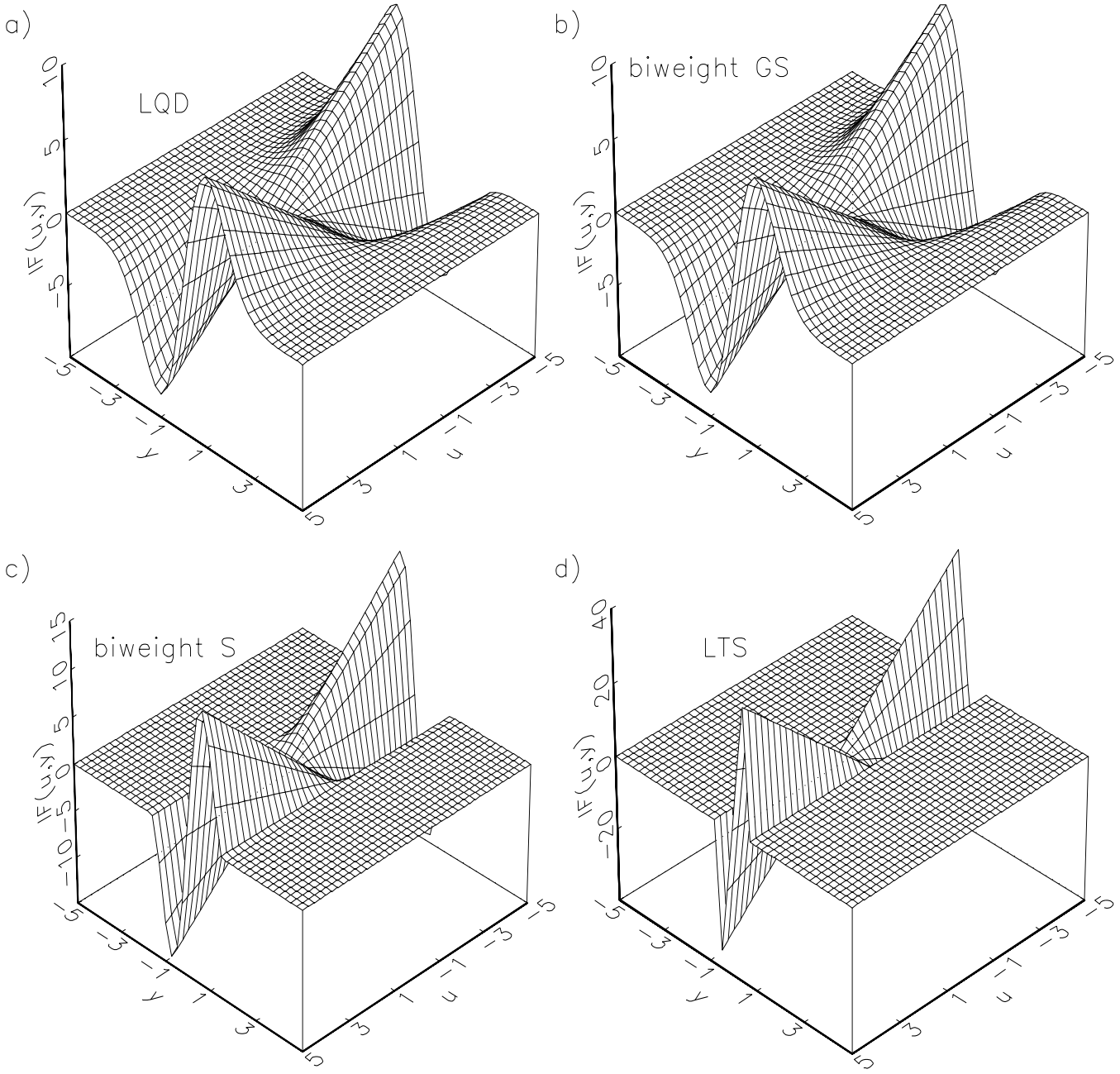


*Figure 3. Influence functions of the LQD, biweight GS-, biweight S-, and LTS estimators.*

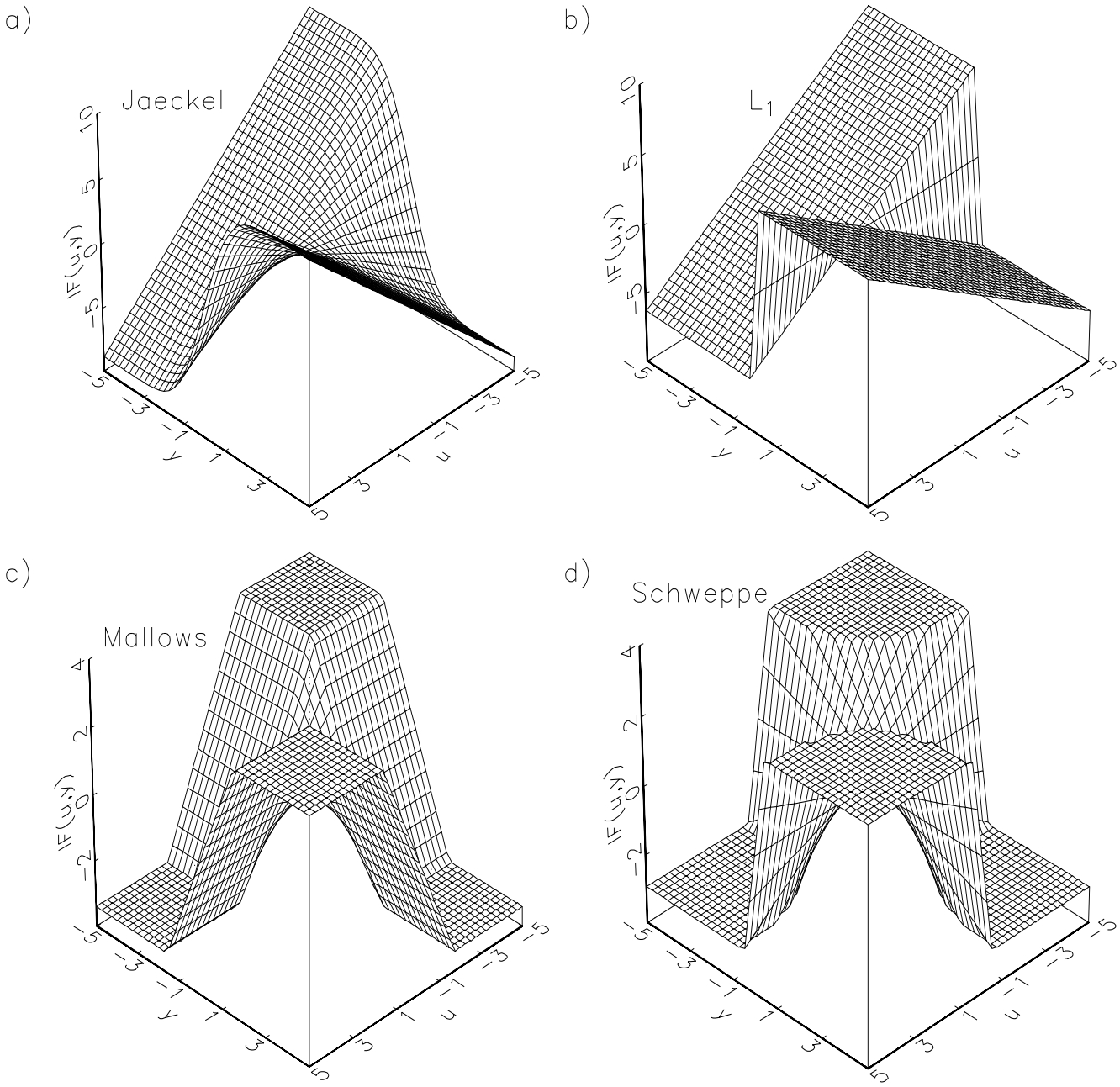


*Figure 4. Influence functions of the Jaeckel estimator based on Wilcoxon scores, the $L_1$ estimator, and the optimal robust 95% efficient Mallows and Schweppe estimators.*

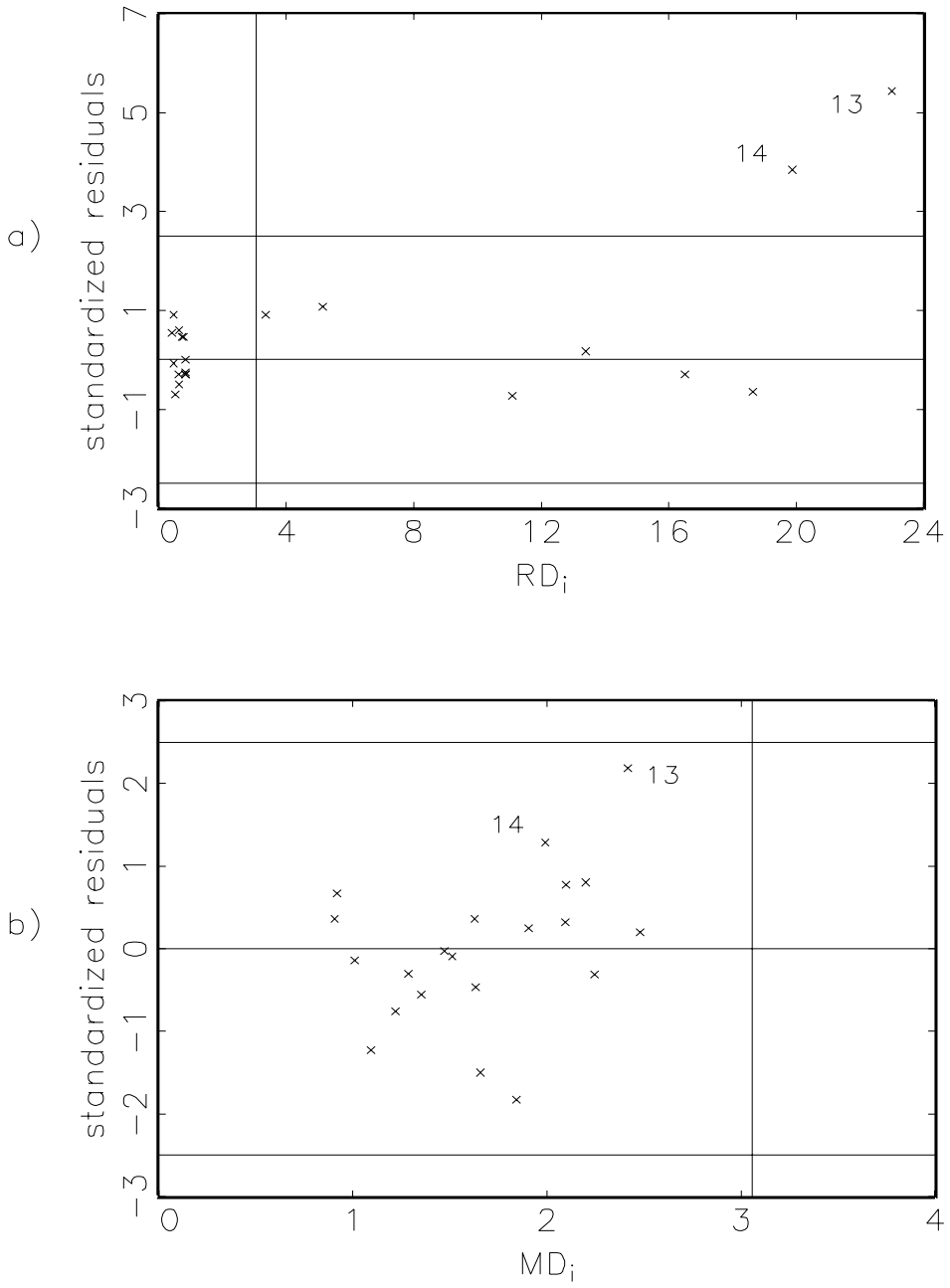


*Figure 5. Diagnostic plots of nitrogen data: a) standardized robust residuals obtained by the LQD method versus robust distances $RD_i$ based on the MVE; b) standardized LS residuals versus Mahalanobis distances.*